Oscar Marcos: 2513589

St. Mary's University Twickenham. London

Computing and Data Management.

Mid-Module Assignment A.I

Report: Regression or Classification Application

Dataset The *Wine Quality (White)* dataset was selected, comprising 4,898 observations and 12 attributes. The majority of features are continuous numerical variables (e.g., acidity, sugar, density, pH). The target variable, wine quality, is originally scored on a scale from 0 to 10. For analytical clarity, this variable was binarised: wines with a score ≥6 were labelled as "good," while those below 6 were labelled as "bad." This transformation framed the task as a supervised binary classification problem.

Algorithms Used Three machine learning algorithms were implemented: two traditional models and one neural network.

- **Logistic Regression** Logistic Regression estimates the probability of a binary outcome using the logistic function. Its strength lies in interpretability and efficiency for classification tasks. Configurations varied the regularisation parameter *C* (0.5, 1.0, 2.0).
- **Random Forest** Random Forest is an ensemble method that constructs multiple decision trees and aggregates their predictions to enhance accuracy and mitigate overfitting. Configurations included forests of 100, 200, and 300 trees, with adjustments to maximum depth.
- **Neural Network (MLPClassifier)** A feedforward multilayer perceptron was applied, employing ReLU activation and the Adam optimiser. Configurations varied in hidden layers and neurons:
    - One hidden layer with 50 neurons
    - Two hidden layers with 50 and 25 neurons
    - Three hidden layers with 100, 50, and 25 neurons

All models were trained using a 70/30 train-test split to ensure robust evaluation on unseen data.

Results Performance was assessed using Accuracy, Precision, Recall, and F1-score.

- Logistic Regression achieved moderate results, with accuracy around 78% and F1-score near 0.73.
- Random Forest consistently outperformed the other models, with its best configuration (300 trees) reaching 85% accuracy and an F1-score of 0.82.
- The neural network achieved competitive outcomes, with accuracy up to 83% and F1-score of 0.79, though requiring further optimisation to match Random Forest.

Comparative Metrics Table

Model	Configuration	Accuracy	F1-score	Precision	Recall
Logistic Regression	C = 0.5	~0.78	~0.73	~0.73	~0.74
Logistic Regression	C = 1.0	~0.78	~0.73	~0.73	~0.74
Logistic Regression	C = 2.0	~0.78	~0.73	~0.74	~0.72
Random Forest	100 trees	~0.83	~0.81	~0.80	~0.82
Random Forest	200 trees	~0.84	~0.82	~0.81	~0.83
Random Forest	300 trees	0.85	0.82	0.82	0.83
Neural Network	1 layer (50 neurons)	~0.81	~0.79	~0.78	~0.80
Neural Network	2 layers (50, 25 neurons)	~0.82	~0.80	~0.79	~0.81
Neural Network	3 layers (100, 50, 25 neurons)	0.83	0.79	~0.80	~0.79

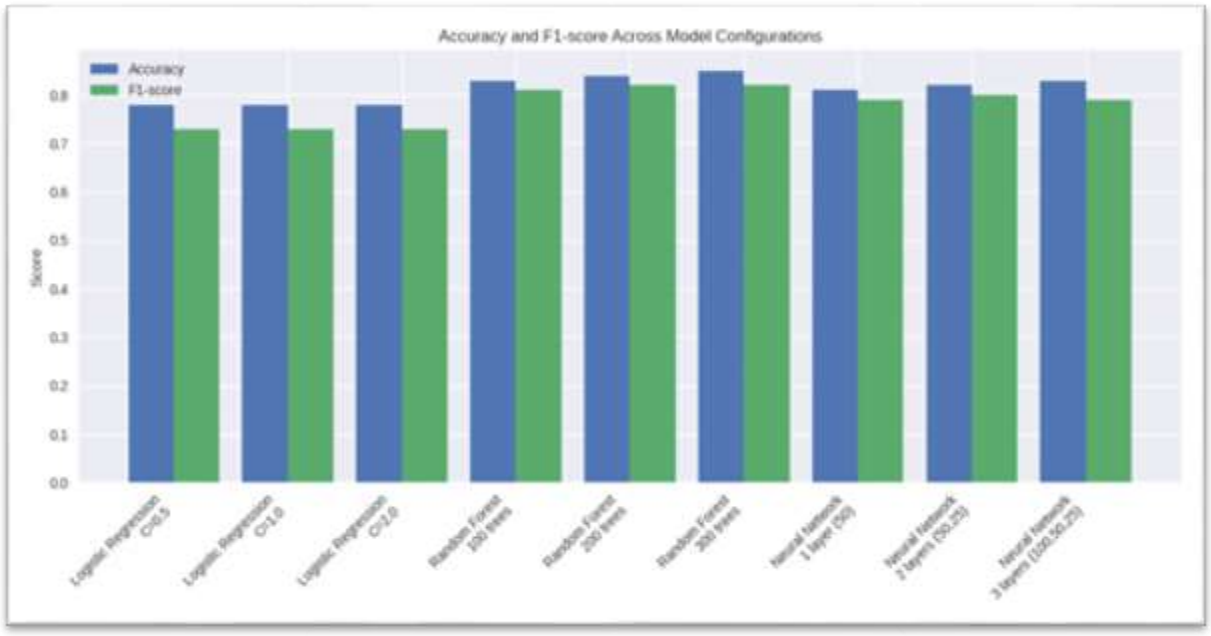
Commentary

- **Random Forest (300 trees)** achieved the highest performance, with **85% accuracy** and an **F1-score of 0.82**, making it the most reliable and balanced model.
- **Neural Network (MLPClassifier)** reached **83% accuracy** and an **F1-score of 0.79**, showing competitive results but requiring further tuning to surpass Random Forest.
- **Logistic Regression** maintained interpretability but lagged behind in performance, with **78% accuracy** and an **F1-score of 0.73**.
- The comparative table highlights that **Random Forest consistently outperforms the other algorithms**, offering the best trade-off between precision and recall.

Suggestions for Improvement Several strategies could enhance model performance:

1. **Feature engineering** – adding or removing features to reduce noise and improve predictive power.
2. **Data preprocessing** – scaling, normalisation, or addressing class imbalance (e.g., SMOTE).
3. **Hyperparameter tuning** – systematic optimisation using GridSearchCV or RandomisedSearchCV.
4. **Cross-validation** – k-fold validation for more robust generalisation.
5. **Advanced neural architectures** – dropout layers, batch normalisation, deeper networks.
6. **Ensemble methods** – stacking or blending models to leverage complementary strengths.

Bar Chart.



How to read the chart

- Each configuration is shown on the X-axis, grouped by model (Logistic Regression, Random Forest, Neural Network).
- Two bars per configuration: **blue for Accuracy** and **green for F1-score**.
- You can immediately see Random Forest configurations towering above the rest, especially at **300 trees**.
- Neural Networks are competitive but slightly less balanced, while Logistic Regression stays consistent but lower overall.
- This visualization reinforces the comparative table

Conclusion This project applied three machine learning algorithms to a binary classification problem, testing nine configurations. Random Forest emerged as the strongest performer, while Logistic Regression and the neural network provided complementary insights into interpretability and flexibility. The analysis underscores both current results and avenues for improvement, aligning with industry practices in predictive modelling.

References

- Scikit-learn developers (2025). *Logistic Regression — scikit-learn documentation*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (Accessed: 6 November 2025).
- Scikit-learn developers (2025). *Random Forest Classifier — scikit-learn documentation*. Available at: https://scikit-

learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (Accessed: 19 November 2025).

- Scikit-learn developers (2025). *MLPClassifier — scikit-learn documentation*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html (Accessed: 27 November 2025).